WATERSHED SIMILARITY ANALYSIS FOR MILITARY APPLICATIONS USING SUPERVISED-UNSUPERVISED ARTIFICIAL NEURAL NETWORKS

B. B. Hsieh* and M. R. Jourdan
US Army Engineer Research and Development Center
3909 Halls Ferry Road, Vicksburg, MS 39180, USA

ABSTRACT

Incorporation of Geographic Information Systems (GIS) into Unsupervised-Supervised Artificial Neural Networks (ANNs) was applied to quantify the similarity of watershed characteristics. The goal of this approach is to find the best match watershed from a large knowledge base of over one thousand quantifying watersheds and to determine the reliability of "transplant" watershed information during the clustering and classification stages. The prediction stage of the study compares the hydrographs between this unknown watershed and the best-selected watershed to verify the similarity performance. Three examples demonstrate use of random selection, average size, and median size watersheds to test the reliability of developing procedures. It is shown that the basin area ratio provides a reasonable conversion factor for adjusting the magnitude of the predictive hydrograph. While the monthly hydrographs comparison receives very satisfactory agreement, the daily hydrographs comparison also obtains reasonable results when a high degree of similarity is found in the knowledge base.

1. INTRODUCTION

It is noted that this is the continuation of system development in the original study (Hsieh, et al. 2004) of watershed similarity. Three new parameters were added and the effort extended from the previous investigation. They were – increasing the number of watersheds from 193 to 1064, adding land use/land cover and soil type parameters, and most importantly, performing the verification (prediction) process.

The ability to predict watershed hydrologic conditions and the associated potential for flooding to occur plays a significant role in planning and operational activities. To make highly accurate hydrologic predictions, either physically-based or system-based, the system parameters and prediction variables are sometimes unavailable or even totally missing. This certainly curtails the capability of prediction, particularly for operations where very little time is available to conduct the analysis. Very often, the information for a particular watershed may be entirely unavailable; this situation could be resolved by the similarity concept.

The purpose of this approach is to find the best match watershed from a large knowledge base and to determine the reliability of "transplant" watershed information such as hydrologic and climatic parameters. The degree of similarity is based on inter and intra relationships among many geologic, soil, hydrologic and climatic factors. Various methods have been employed to analyze the similarity between two objects.

2. BASIC CONCEPTS OF ANNs

ANNs (Artificial Neural Networks) are one of many emerging computing technologies that have been actively studied over last three decades (Hechet-Nielsen, 1990). They are inspired by ideas from neuroscience that a sophisticated computing system can be constructed from a simple processing unit. How neural networks work depends on the interconnectivity between neurons. An artificial neuron itself carries out very simple signal processing using its internal function, which is usually a nonlinear function such as a sigmoid function. Due to this nonlinear nature of neurons, a massively connected network of neurons can capture very complex and highly nonlinear characteristics of data (Fayyad, 1996). When neural networks are used to capture complex structural information of the feature space, it is often necessary to analyze what the networks have learned or discovered in addition to just using them to obtain answers for unknown input data (Bigus, 1996).

Scientific and engineering communities have reported ANNs theoretical development and applications for several decades, particularly for supervised ANNs. In this paper, a brief description of the unsupervised ANNs and the concept of the integration of unsupervised-supervised ANNs are discussed. Learning or adaptation, in which a desired response can be used by the system to guide the learning process, is called supervised learning, while unsupervised learning is learning in which the system parameters are adapted using only the information of the input and constrained by prespecified internal rules.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **01 NOV 2006**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
Watershed Similarity Analysis For Military Applications Using Supervised-Unsupervised Artificial Neural Networks

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Engineer Research and Development Center 3909 Halls Ferry Road, Vicksburg, MS 39180, USA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002075., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

3. SUPERVISED AND UNSUPERVISED ANNs

The vast majority of artificial neural network solutions have been trained with supervision. In this mode, the actual output of a neural network is compared to the desired output. The network then adjusts weights, which are usually randomly set to begin with, so that the next iteration, or cycle, will produce a closer match between the desired and the actual output. The learning method tries to minimize the current errors of all processing elements. This global error reduction is created over time by continuously modifying the input weights until acceptable network accuracy is reached.

With supervised learning, the artificial neural network must be trained before it becomes useful. Training consists of presenting input and output data to the network. This data is often referred to as the training set. That is, for each input set provided to the system, the corresponding desired output set is provided as well. In most applications, actual data must be used. This training phase can consume a lot of time. In prototype systems, with inadequate processing power, learning can take weeks. This training is considered complete when the neural network reaches a user defined performance level. This level signifies that the network has achieved the desired statistical accuracy as it produces the required outputs for a given sequence of inputs. When no further learning is necessary, the weights are typically frozen for the application. Some network types allow continual training, at a much slower rate, while in operation. This helps a network to adapt to gradually changing conditions.

Unsupervised training is when the networks learn from their own classification of the training data, without external help. It is assumed that class membership is broadly defined by the input patterns sharing common features, and that the network will be able to identify those features across the range of input patterns.

Unsupervised learning is the great promise of the future. It shouts that computers could someday learn on their own in a true robotic sense. Currently, this learning method is limited to networks known as self-organizing maps. These kinds of networks are not in widespread use. They are basically an academic novelty. Yet, they have shown they can provide a solution in a few instances, proving that their promise is not groundless. They have been proven to be more effective than many algorithmic techniques for numerical aerodynamic flow calculations. They are also being used in the lab where they are split into a front-end network that recognizes short, phoneme-like fragments of speech, which are then passed on to a back-end network. The second artificial network recognizes these strings of fragments as words.

4. AN UNSUPERVISED ANNs (SOFM)

Self-organizing Feature Maps (SOFM) is a special kind of neural network that can be used for clustering tasks. Only one map node (winner) at a time is activated corresponding to each input. The location of the responses in the array tends to become ordered in the learning process as if some meaningful nonlinear coordinate system for the different input features were being created over the network. This illustrates an important and attractive feature of SOFM applications, in that a multi-dimensional input ensemble is mapped into a (one or) two-dimensional space, preserving the topological structure as much as possible. Boogaard, Ali, and Mynett (1998) applied the SOFM to hydrological and ecological data sets.

The SOFM is trained without teacher signals (unsupervised), unlike some other ANNs in which supervised training is used, as in backprobagation networks. The learning algorithm used in this study is the same as Kohonen's algorithm (Kohonen, 1989). SOFM is a special kind of neural network that can be used for clustering tasks. Only one map node (winner) at a time is activated corresponding to each input. The location of the responses in the array tends to become ordered in the learning process as if some meaningful nonlinear coordinate system for the different input features were being created over the network. This illustrates an important and attractive feature of SOFM applications, in that a multi-dimensional input ensemble is mapped into a (one or) two-dimensional space, preserving the topological structure as much as possible. Boogaard, Ali, and Mynett (1998) applied the SOFM to hydrological and ecological data sets.

The SOFM is a set of artificial neurons, which are ordered in N^n space. A two dimensional array (n=2) is the most common map and is used to map an input signal in R^m (m >n) space onto the two-dimensional space. Basically, an SOFM typically consists of two layers. One is an input layer into which input feature vectors will be fed and other layer is a two-dimensional competitive layer, which orders the neuron's responses spatially. Neurons can be arranged on a rectangular map so that they can be implemented using a simple 2-D data array. A hexagonally arranged neuron map is, however, often used because it has the advantage of the Euclidean distance (equi-distance) between adjacent neurons. (Kohonen, 1995).

5. VISUALINZING A SOFM

Visualization techniques to depict the data structure of the feature space in the form of clustering of neurons in the 2-D SOFM have been developed (Ultsch, 1993). This visualization typically uses the grayscale to illustrate the distance between connection weights. The light shading typically represents a small distance and the dark shading represents a large distance

This type of visualization is useful as long as relatively clear cluster boundaries exist or the granularity of the distance differences is large. When the cluster boundaries get fuzzy or the granularity of the distances become too small to represent with the grayscale, it becomes increasingly difficult to identify fuzzy cluster landscapes. Moreover, since all distance values are normalized, only relative (qualitative) analysis is allowed. Subsequently, this “grayscale distance map” cannot be used to compare different SOFM mapping results.

When the SOFM is used to discover some structure of the given samples in the feature space, it is often useful to visualize the finding in the form of clustering formations. Visualization techniques to depict the data structure of the feature space in the form of clustering of neurons in the 2-D SOFM have been developed (Ultsch, 1993). This visualization typically uses the grayscale to illustrate the distance between connection weights. The light shading typically represents a small distance and the dark shading represents a large distance.

NeuroDimensions (2001) developed a visual version for the Kohonen topological feature maps to check the performance of SOFM. Three basic windows used for evaluating the clustering are quantization metric, united distance, and frequency.

Quantization Metric: It produces the average quantization error, which measures the goodness of fit of a clustering algorithm. It is the average distance between each input and the winning process element (PE). If the quantization error is large, then the winning PE is not a good representation of the input. If it is small, then the input is very close to the winning PE. The quantization error is best for comparing the clustering capabilities between multiple trainings of the same SOFM on the same point.

Unified Distance: This is the distance between PE clustering centers. The weights from the input to each PE cluster centers of the SOFM. Inside a cluster of inputs, SOFM PEs will be close to each other.

Frequency: Typically, the number of SOFM PEs is much larger than the number of clusters expected. This allows multiple PEs to capture one logical cluster. The SOFM map is a group of PEs representing a single cluster of the input.

6. GIS LINKED TO ANNs

GIS data often includes satellite and other remotely sensed imagery. An example of the analysis of imagery involves either supervised or unsupervised classification. Unsupervised classification of imagery involves the analysis of color or black and white pixels of the image for the purposes of classifying image objects and entities, where, tone, texture and hue are used. Supervised classification of imagery involves referencing the pixels to actual field or site conditions and color balancing of the image for similar classification purposes.

ANNs are increasingly being used for the purpose of determining spatial patterns. In the area of landscape ecology, the landscape pattern is an important factor enabling classification. Indeed, more recent developments in the area of remote sensing analysis involve ANNs for the analysis of images for the purposes of classifying objects.

7. GEOSPATIAL KNOWLEDGE BASE DEVELOPMENT

Geospatial data of geographic locations and characteristic natural and constructed features were gathered for the database development. GIS databases were utilized for this endeavor; specifically the EPA’s Better Assessment Science Integration Point & Non Point Sources (BASINS) system provided the 300-meter USGS Digital Elevation Model (DEM), Land Use/Land Cover, Soils (STATSGO), and watershed gauge locations within the conterminous United States of America (Figure 2). These gauge locations were selected with the criteria of 100 percent complete dataset for medium- to moderately large- sized basins, 6 to 7900 km^2.

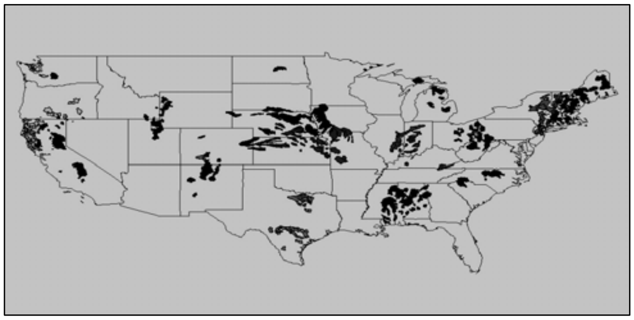


Figure 1. Watersheds within the contiguous United States

Watershed development was conducted with the Environmental Systems Research Institute's ArcGis/ArcView and the Department of Defense's Watershed Modeling System (WMS). From the GIS databases, data was extracted, projected and shaped into Arc/Info griddled ASCII data as input into the WMS interface where basin delineation and parameter estimations were conducted. Watershed parameters such as Drainage Area, Basin Slope, Basin Length, Basin Perimeter, etc. were among the variables derived for the ANN's analyses. Watersheds selected were within a 10 percent margin of error when the areas were compared with recorded drainage areas from BASINS.

From these selected watersheds, mean daily flow data for their respective periods of record were compiled for the ANN's' verification process. In addition, thirty-year mean monthly and annual precipitation, as well as temperature data, were derived from *PRISM (**P**arameter-elevation **R**egressions on **I**ndependent **S**lopes **M**odel)* and presented as GIS coverages. Subsequent GIS analyses produced mean monthly and annual, precipitation and temperature data, for all selected basins. The final knowledge base has the dimension of a 1064 watersheds x 70 variables matrix with final relevant parameters listed as follows.

Geometric Parameters:

Basin Area, Basin Slope, Basin Average Elevation, Basin Shape Factor, Basin Sinuosity Factor, Average Overland Flow Distance, Maximum Flow Distance, Maximum Flow Slope, Maximum Stream Slope, Centroid to Nearest Point of MaxFlowDist

Land Use/Land Cover Parameters:

Residential/Industrial, Agricultural Land, Rangeland, Forest Land, Open Water, Wetlands, Exposed Rock, Tundra, Glaciers

Soil Type Parameters:

Sands and Gravel, Silts, Sandy Loam, Clays

Hydrologic Parameters:

Seasonal and Annual Precipitation, Seasonal and Annual Temperature

A data-driven computational procedure including knowledge base and two components of ANNs (clustering and classification) and prediction (verification) was developed (Figure 2). Takatsuka (2002) applied SOFM and interactive 3-D visualization to geospatial data. Ultsch, Korus, and Kleine (1995) developed the integration of neural networks and knowledge-based systems in medicine.

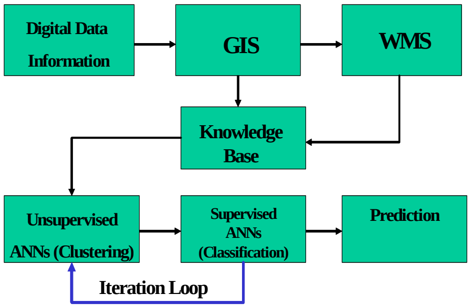


Figure 2. System components for similarity analysis

8. DEMONSTRATION EXAMPLES

From the knowledge base, all the geometric parameters, the land use/land cover, the soil types and the seasonal and annual mean values of both precipitation and temperature were used to test this calculation procedure. In order to test the reliability of the system development, three sizes of watershed are selected to examine the performance. The detailed search process is only presented in the first example.

8.1 Random selection (watershed 4288000)

The goal of this test is to use a known watershed (gage number 4288000) to search for the best similar watershed. This part of study is divided into two portions. While the clustering analysis is used to identify the similarity between the watersheds, the classification analysis is used to verify the clustering performance. To check the reliability of the prediction, time series hydrographs are used to compare the resulting search pattern. In this procedure, the hydrograph of gage 4288000 is hidden purposely in order to check the performance of the system once the best similar watershed is found.

During the clustering computation, a 5 x 5 matrix of SOFM is initially selected. Through repeated iterations (usually 200) of the examination of frequency, unified distance, and quantization of the unsupervised synapse, an optimal clustering set to distribute the winner for each watershed is obtained (Figure 3). The numbers in this 5 X 5 matrix show the most similar watershed within the same group (there are 25 groups in this case).

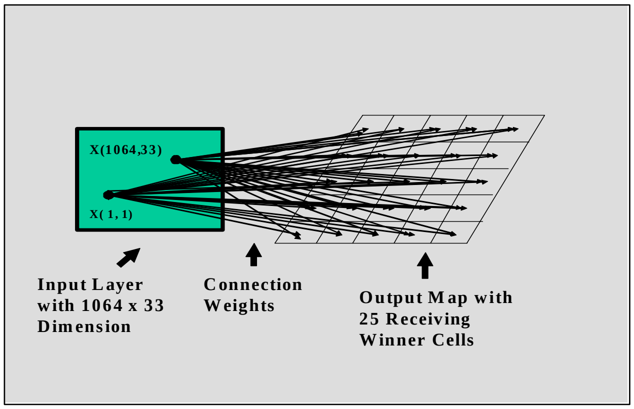

X(1064,33)
X(1, 1)
Input Layer with 1064 x 33 Dimension
Connection Weights
Output Map with 25 Receiving Winner Cells

Figure 3. SOFM with 5 x 5 matrix lattice

For classification, the problem was trained with (Multi-layered Feed Forward neural Networks (MLP) ANNs and the outcome provided the confidence level of the clustering analysis, which resulted in a successful classification rate of about 91 percent meeting the target. This result indicates that watershed 4288000 belongs to the group with 103 (group 7) most similar watersheds from the original 1063 possible candidates. This clustering-classification process is repeated until the final target watershed is found. This process is called search iteration. Figure 4-5 show the classification verification during the first and second search iterations respectively. It is noted that the size of clustering for this iteration has been reduced to a 3 X 3 matrix.

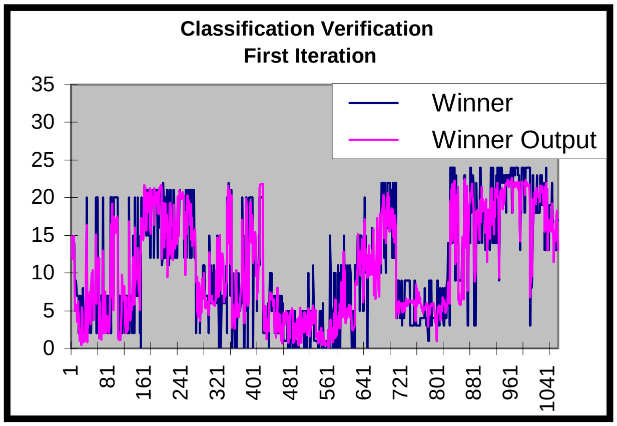


Figure 4. Classification verification for the first system iteration with 1064 watersheds (group – X-axis; number of assigned watershed – Y-axis)

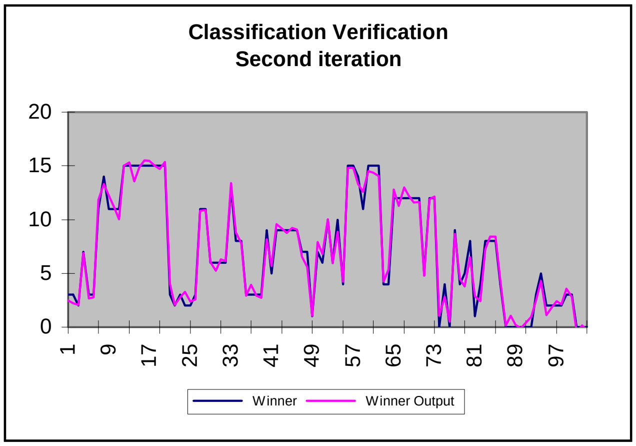


Figure 5. Classification verification for the second system iteration with 103 watersheds (group – X-axis; number of assigned watershed – Y-axis)

The final candidate for this search is the watershed number 01144000. This implies that the flow patterns from station 01144000 will be most similar to those of station 04288000. Flow hydrograph comparisons between these two stations during the period 1999-2001 are shown in Figure 6. Although the flow pattern, particularly, the phase matches very well, the performance of the amplitude representations is dissatisfactory.

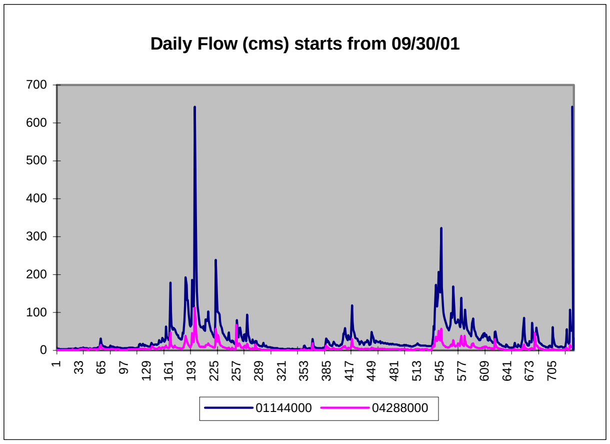


Figure 6. Most similar flow (cms) (01144000 - pink) vs. observed flow (04288000 - blue) flow – X-axis; days – Y-axis)

When examining the involved parameters between these two watersheds, the area and maximum flow distance showed a significant difference. The estimated hydrograph was adjusted by taking the area ratio of station 04288000 and station 01144000 (Figure 7).

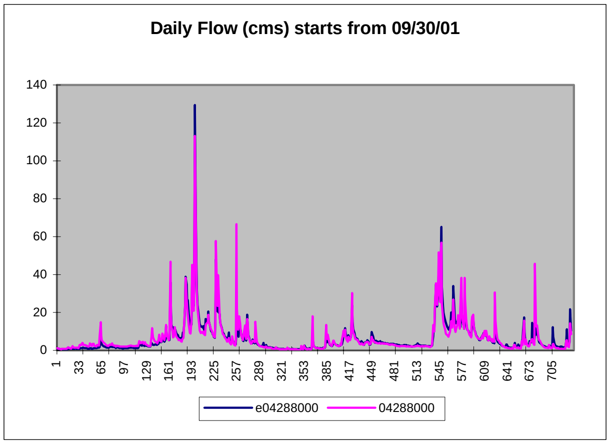


Figure 7. Flow (cms) estimation (e04288000 - pink) vs. observed flow (04288000 - blue) after basin area ratio adjustment for 33 inputs approach (flow – X-axis; days – Y- axis)

The major element in making this integration system a success is to tune the clustering group as well as rechecking the performance of the classification process. But the identification of the reliability for application also requires data on how well the "transplant" performs. Therefore, a series of combinations including the features of input parameters is adopted. Table 1 summarizes the performance due to the selection of input parameters. This indicates that the important group parameters are hydrologic, geometry, soil type, and land use. The performance difference between geometry and hydrologic groups is quite small. The magnitude of hydrographs could not be adjusted by ratios obtained using hydrologic, soil type, and land use groups since they do not contain the basin area factor after the best candidate is found.

Table 1. Sensitivity test due to input parameters

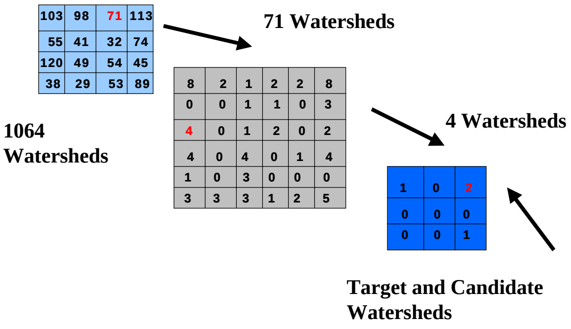
Parameters	Candidate Watershed	Correlation Coefficient	Mean Error
All Groups	0114000	0.92	0.17
Geometry	4282000	0.82	-7.80
Hydrologic	4288000	0.83	2.99
Land Use	2472000	0.10	-28.34
Soil Type	1170100	0.67	- 4.25

8.2 Average size (watershed 11427000)

This demonstration example uses an average size watershed (856.97 square kilometers) to perform the same search procedure as the first example. Instead of using short-term hydrographs for comparison of results, it uses much a longer period to compare the daily and monthly flow conditions. In addition, a number of statistical parameters are computed to check the degree of similarity between this target watershed and the best candidate watershed.

As in the first demonstration example, the clustering-classification iteration processes were conducted until the best similar watershed of target watershed from the knowledge base is found. This time we use a 3x3, 6x6, and 3x3 clustering sequences approach to find the best candidate. Figure 8 summaries this approach progressively.

The watershed number 1144500 is the final candidate to this search process. They fall into the same group after three clustering-classification iterations. To examine the similarity, the comparisons of daily and monthly flow for 34 years (1954-1987) between target and best candidate watersheds are shown in Figure 9-10. While the phase comparison receives good results, the amplitude remains underestimated results after the area ratio factor was applied. This could indicate that more watersheds need to be included in the knowledge base and that the area ratio factor may not be the only function useful in final conversion.



103	98	71	113
55	41	32	74
120	49	54	45
38	29	53	89

8	2	1	2	2	8
0	0	1	1	0	3
4	0	1	2	0	2
4	0	4	0	1	4
1	0	3	0	0	0
3	3	3	1	2	5

1	0	2
0	0	0
0	0	1

Figure 8. A clustering sequences to search the best similar watershed from a given target watershed

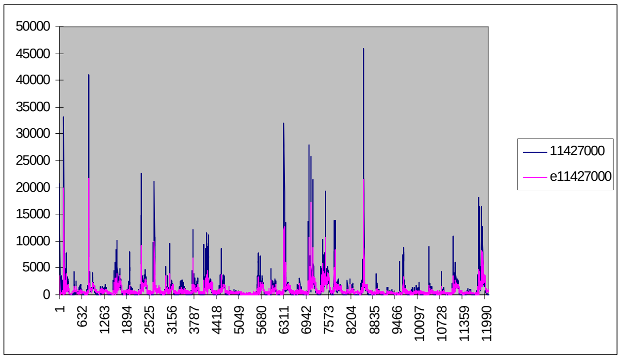


Figure 9. 34-years daily flow (cfs) estimation (pink) versus observation (blue) from watershed 11445500; r=0.883

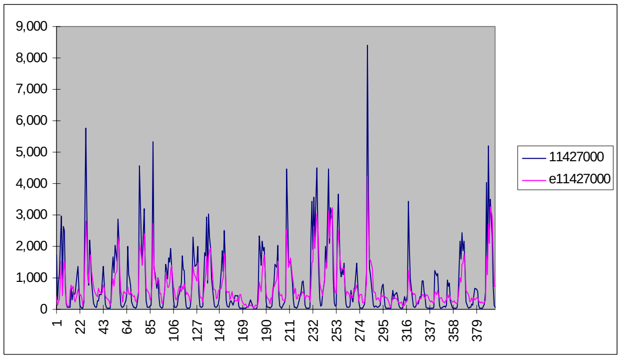


Figure 10. 34-years monthly flow (cfs) estimation (pink) versus observation (blue) from watershed 11445500; r=0.882

Fourteen statistical parameters from true and estimate monthly flow are summarized as Table 2. The deviations between these parameters show that some improvements need to be incorporated into the system.

Table 2. Statistical parameters comparison for 34-years monthly flow (cfs) for target watershed

Watershed Number	11427000	11427000 (estimate)
Mean	825	741
Standard Error	54	33
Median	391	490
Mode	81	969
Standard Deviation	1091	657
Sample Variance	1190503	432541
Kurtosis	8.39	4.12
Skewness	2.40	1.93
Range	8389	4196
Minimum	13.4	44.1
Maximum	8403	4240
Sum	326908	293551
Count	396	396
Confidence Level (95%)	107.79	64.97

8.3 Median size (watershed 03153000)

The target watershed 03153000 with a basin area of 419.6 square kilometers is used to represent another testing case. It also requires three clustering-classification iterations to achieve the final search. The best candidate watershed for the similarity analysis is the watershed 03152000, which has the basin area 1000.2 square kilometers.

Although 9-years (1967-1975) of daily flow prediction shows some overestimation, particularly for the peak flow conditions; the monthly flow prediction receives very good agreement. Figure 11 shows the daily flow comparison for the target watershed, while Figure 12 illustrates the results of the monthly flow prediction.

The frequency analysis (Figure 13) shows the frequency distribution comparison for the estimated and true monthly flow. It is noted that each unit of X-axis represents 100 cfs. Except when the flow volume is around 200 cfs, the monthly distribution reaches very nice agreement.

CONCLUSIONS

An integration of database and ANNs learning was used to identify a very complex nonlinear watershed similarity analysis for military hydrology applications. While the unsupervised ANNs, such as SOFM, were used to perform the clustering of watershed characteristics, the supervised ANNs were used to identify the best match candidate watershed for classification analysis. The search procedure requires several iterations of the clustering-classification loop. The current knowledge base consists of 67 geometric, hydrological, land use, and soil type factors for 1064 selected watersheds. After removing the internal dependency and examining the annual and season representation, 33 factors were selected for final analysis.

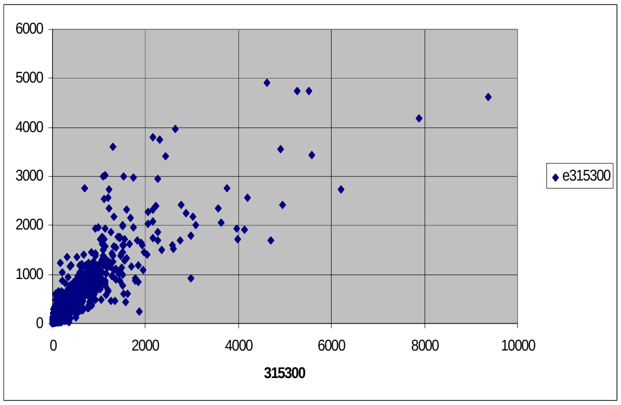


Figure 11. 9-years daily flow (cfs) comparison for median target watershed 03153000 with r=0.896

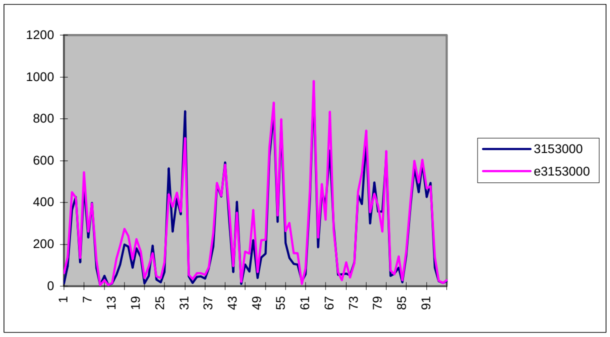


Figure 13. 9-years monthly flow (cfs) comparison for median target watershed 03153000 (blue – observation; pink – estimation) with r=0.976

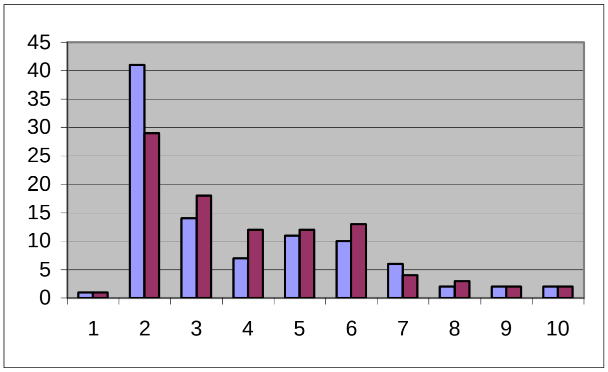


Figure 14. Frequency distribution between true (blue) and estimate (red) monthly flow (x 100 cfs) for target watershed 03153000

Three demonstration examples, including random selection, average size, and median size watersheds were used as the target to search for the best match corresponding candidate. The first example obtained a good correlation coefficient (0.92) for hydrograph prediction (2 years daily flow). It is found that the basin area ratio provides a reasonable factor for making the adjustment for hydrograph prediction. The preliminary sensitivity tests indicate that the hydrologic factors are the best factors in producing a fitness for transplant. In general, monthly hydrograph comparisons have better agreement than the daily hydrographs for both average and median size watershed examples. The most significant reliability is obtained when many watershed patterns are included in the knowledge base. Development of an automated search procedure for a unique solution is the direction proposed for further research.

ACKNOWLEDEMENT

The Military Hydrology Research Program in the Coastal and Hydraulics Laboratory at the US Army Engineer Research and Development Center funded this effort. Permission was granted by the Chief of Engineers to publish this information. Ms. Moira Fong, Miss Woodman P. Wardlaw and Ms. Amy Rainer contributed their efforts in the development of the knowledge base. Their efforts are gratefully appreciated.

REFERENCES

Boogaard, H., Ali Md. S., and Mynett A. E., Self-organizing feature maps for the analysis of hydrological and ecological data set, Proc. of 3rd International Conference on Hydroinformatics, Demark, (1998), pp 733-740.

Hsieh, B., B., Fong, M., T., Jorgeson, J., D., and Skahill, B., E.,Watershed similarity analysis using integration of GIS and unsupervised-supervised ANNs, *Proc. of* 6th International Conference on Hydroinformatics, Singapore, (2004).

NeuroSolutions v4.2, Developers level for windows, NeuroDimension, Inc., Gainesville, FL, (2001).

Takatsuka, M., An application of the self-organizing map and interactive 3-D visualization to geospatial data (from internet), (2001), pp 1-9.

Ultsch, A., Korus, D., and Kleine, T. O., Integration of neural networks and knowledge-based system in medicine, 5th Conference on Artificial Intelligence *in Medicine Europe AIME'95*, Italy, (1995), pp 425-426.